STATS 4CI3/6CI3 Winter 2021

ASSIGNMENT 1

Submit to Crowdmark using the link that was emailed to you.

Due before 11 PM on Wednesday, January 27th.

Your assignment must conform to the Assignment Standards listed below.

Assignments submitted up to 24 hours late will incur a 30% penalty. Assignments submitted more than 24 hours late will receive a zero grade.

Answer all questions, stating your answer and showing your code. Not all questions carry equal marks.

1. **(30 MARKS)** Answer the following, showing all code and R output:

   (a) Produce a vector "v" containing the integers from -15 to 15.

   (b) Use the function "seq" to reproduce the following sequences:

   - v1 $= 1, 2, 3, 4, 5$
   - v2 $= 3, 6, 9, 12, 15, 18, 21, 24, 27, 30$

   (c) Use the "which" function to determine which elements of v2 are greater than 10.

   (d) Use the "curve" function to plot the following curve: $sin(x)$ for $0 < x < 8\pi$.

   (e) Construct a "while" loop to compute the sum of squares from 1 to 10.

2. **(30 MARKS)** Go to the following website,"UCI Machine Learning Repository" and select a dataset that is suitable for visual data analysis. **Do not use a dataset that has been used in class.**

   (a) Report the name of your dataset and the URL that you downloaded it from.

   (b) Produce a summary table (this can be typed up based on R output) for your dataset including variable names, number of observations, etc.

   (c) Using your dataset, produce the following **2** graphs, with labelled axes and Titles: (you do not have to use all of the variables in the production of the graphs)

      i. a histogram or "bar chart"

      ii. a scatter plot

   (d) Create an image that displays both graphs, side by side, in one image.

3. **(20 MARKS)** The null hypothesis is that the mean of the x population is equal to the mean of the y population: $H_0 : \mu_x = \mu_y$.

   Set your seed to 6573. Simulate a sample of 50 observations from a normal distribution with mean 80 and standard deviation 20 using the "rnorm" function and store the vector of values in the variable x. Then simulate a sample of y's by simulating 50 values from an $\mathrm{N}(80, 20)$ distribution and store these values in the variable y. With reference to a similar example that we did in class, answer the following questions: (when a significance level is required, use $\alpha = .05$)

   (a) Calculate the test statistic.

   (b) Calculate the critical value.

   (c) Do you reject the null hypothesis? Why?

4. **(20 MARKS)** The null hypothesis of a lower tail test of the population mean can be expressed as follows: $\mu_0 \leq \mu$, where μ_0 is a hypothesized lower bound of the true population mean μ.

   Define the test statistic z in terms of the sample mean $\bar{x}$, the sample size n, the hypothesized value of the mean μ_0 and the population standard deviation σ:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

   Suppose the manufacturer of batteries **claims that the mean lifetime of a battery is more than 8,000 hours**. In a sample of 50 batteries, it was found that they only last 7900 hours on average. Assume the population standard deviation is 100 hours. Use a significance level $\alpha = .05$.

   Answer the following questions:

   (a) Calculate the test statistic.

   (b) For this type of one-sided test, the critical value (cv) is defined as follows in terms of R code: cv = qnorm(1 − alpha)
   Calculate the critical value.

   (c) Do you reject the null hypothesis? Why?

Assignment Standards

- LaTeX is strongly recommended but not strictly required. The use of Markdown in R studio is also recommended.
- Submit your assignment as one **.pdf document**. **All R code should be included inline.**
- Do not include a title page. The title, your **name and student number** should be printed at the top of the first page.
- The writing and referencing should be appropriate to the university level.
- Various tools, including publicly available internet tools, may be used by the instructor to check the originality of submitted work.